Docket No.: PF-0066-2 DIV
USSN: 09/225,080
Ref. Tab D

Human cellular protein patterns and their link to genome DNA sequence data: usefulness of two-dimensional gel electrophoresis and microsequencing

JULIO E. CELIS,*,[1] HANNE H. RASMUSSEN,* HENRIK LEFFERS,* PEDER MADSEN,* BENT HONORÉ,* BORBALA GESSER,* KURT DEJGAARD,* JOËL VANDEKERCKHOVE†

*Institute of Medical Biochemistry and Human Genome Research Centre, Aarhus University, DK-8000 Aarhus, Denmark and †Laboratorium voor Fysiologische Chemie, Rijksuniversiteit Gent, Belgium

ABSTRACT **Analysis of cellular protein patterns by computer-aided 2-dimensional gel electrophoresis together with recent advances in protein sequence analysis have made possible the establishment of comprehensive 2-dimensional gel protein databases that may link protein and DNA information and that offer a global approach to the study of the cell. Using the integrated approach offered by 2-dimensional gel protein databases it is now possible to reveal phenotype specific protein (or proteins), to microsequence them, to search for homology with previously identified proteins, to clone the cDNAs, to assign partial protein sequence to genes for which the full DNA sequence and the chromosome location is known, and to study the regulatory properties and function of groups of proteins that are coordinately expressed in a given biological process. Human 2-dimensional gel protein databases are becoming increasingly important in view of the concerted effort to map and sequence the entire genome.—Celis, J. E.; Rasmussen, H. H.; Leffers, H.; Madsen. P.; Honoré, B.; Gesser, B.; Dejgaard, K.; Vandekerckhove, J. Human cellular protein patterns and their link to genome DNA sequence data: usefulness of two-dimensional gel electrophoresis and microsequencing. *FASEB J.* 5: 2200–2208; 1991.**

Key Words: human protein patterns · 2-dimensional gel protein databases · gene expression · microsequencing · cDNA cloning · linking protein and DNA information · genome mapping and sequencing

PROTEINS SYNTHESIZED FROM information contained in the DNA orchestrate most cellular functions. The total number of proteins synthesized by a typical human cell is unknown although current estimates range from 3000 to 6000. Of these, as many as 70% may perform household functions and are expected to be shared by all cell types irrespective of their origin. There are many different cell types in the human body with perhaps 30,000 to 50,000 proteins expressed in the organism as a whole judged from the fact that about 3% of the haploid genome correspond to genes. Today only a small fraction of the total set of proteins has been identified, and little is known about the protein patterns of individual cell types or their variation under physiological and abnormal conditions.

For the past 15 years, high resolution 2-dimensional gel electrophoresis has been the technique of choice to determine the protein composition of a given cell type and for monitoring changes in gene activity through quantitative and qualitative analysis of the thousands of proteins that orchestrate various cellular functions (refs 1–6 and references therein). The technique originally described by O'Farrell separates proteins in terms of their isoelectric point (pI) and molecular weight. Usually one chooses a condition of interest and the cell reveals the global protein behavioral response as all detected proteins can be analyzed both qualitatively and quantitatively in relation to each other. At present, most available 2-dimensional gel techniques (regular gel format) can resolve between 1000 and 2000 proteins from a given mammalian cell type, a number that corresponds to about 2 million base pairs of coded DNA. Less abundant proteins can be detected by analyzing partially purified cellular fractions.

Two-dimensional gel ectrophoresis has been widely applied to analysis of cellular protein patterns from bacteria to mammalian cells (refs 1–6, and references therein). In spite of much work, however, information gathered from these studies has not reached the scientific community in its fullness because of lack of standardized gel systems and the lack of means for storing and communicating protein information. Only recently, because of the development of appropriate computer software (7–13), has it been possible to scan gels, assign numbers to individual proteins, and store the wealth of information in quantitative and qualitative comprehensive 2-dimensional gel protein databases (4, 14–23), i.e., those containing information about the various properties (physical, chemical, biological, biochemical, physiological, genetic, immunological, architectural, etc.) of all the proteins that can be detected in a given cell type. Such integrated 2-dimensional gel protein databases offer an easy and standardized medium in which to store and communicate protein information and provide a unique framework in which to focus a multidisciplinary approach to study the cell. Once a protein is identified in the database, all of the information accumulated can be easily retrieved and made available to the researcher. In the long run, protein databases are expected to foster a wide variety of biological information that may be instrumental to researchers working in many areas of biology—among others, cancer and oncogene studies, differentiation, development, drug development and testing, genetic variation, and diagnosis of genetic and clinical diseases (**Fig. 1**).

The approach using systematic 2-dimensional gel protein analysis has recently gained a new dimension with the advent of techniques to microsequence major proteins recorded

[1]To whom correspondence should be addressed, at: Institute of Medical Biochemistry and Human Genome Research Centre, Ole Worms Alle, Bldg. 170, University Park, DK-8000 Aarhus C, Denmark.

0892-6638/91/0005-2200/$01.50.

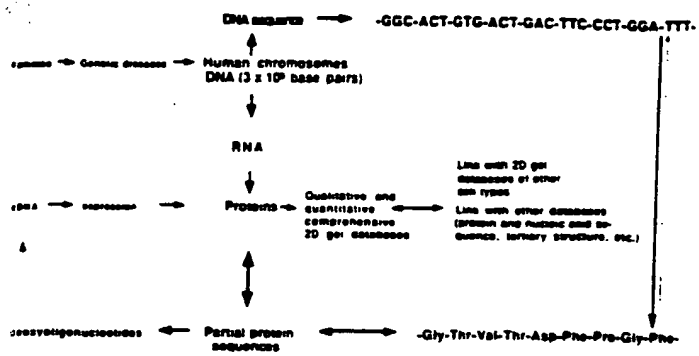


Figure 1. Interface between partial protein sequence databases, comprehensive 2-dimensional gel databases, and the human genome sequencing project. Appropriate software is required to compare protein and DNA sequences. In general, although the inference of a protein's sequence from the DNA sequence (thick arrow) is direct and unambiguous, the DNA sequence can only be inferred approximately from the protein sequence (thin arrow) and cloning of the gene requires either a cDNA or the requisite group of oligonucleotide probes deduced from the partial amino acid sequence. Modified from ref 6.

in the databases (refs 24–42 and references therein). Partial protein sequences can be used to search for protein identity as well as to prepare specific DNA probes for cloning as-yet-uncharacterized proteins (Fig. 1). As these sequences can be stored in the database (see for example Fig. 2H), they offer a unique opportunity to link information on proteins with the existing or forthcoming DNA sequence data on the human genome (Fig. 1) (20, 36, 39).

Using the integrated approach offered by comprehensive 2-dimensional gel databases (Fig. 1), it will be possible to identify phenotype-specific proteins; microsequence them and store the information in the database; search for homology with previously characterized proteins; clone the cDNAs, assign partial protein sequences to genes for which the full DNA sequence and the chromosome location are known, and study the regulatory properties and function of groups of proteins (pathways, organelles, etc.) that are coordinately expressed in a given biological process. Comprehensive 2-dimensional gel protein databases will depict an integrated picture of the expression levels and properties of the thousands of protein components of organelles, pathways, and cytoskeletal systems in both physiological and abnormal conditions and are expected to lead to identification of new regulatory networks in different cell types and organisms. In the future, 2-dimensional gel protein databases may be linked to each other as well as to national and international specialized databanks on nucleic acid and protein sequences, protein structures, NMR experimental data, complex carbohydrates, etc.

A few 2-dimensional gel protein databases that are accessible in a computer form have been published in extenso: these correspond to the protein-gene database of *Escherichia coli* K-12 developed by Neidhardt and colleagues (14, 23), the rat REF 52 database established by Garrels and co-workers at Cold Spring Harbor (18, 22), and a few human databases (transformed amnion cells [15, 20], normal embryonal lung MRC-5 fibroblasts [17, 21], keratinocytes [19] and peripheral blood mononuclear cells [15]) developed in Aarhus. Given space limitations and to keep this review in focus, we will concentrate on the computerized analysis of human cellular 2-dimensional gel patterns, and in particular on the steps involved in establishing comprehensive 2-dimensional gel databases that can link protein and DNA information.

MAKING AND MANAGING A COMPREHENSIVE 2-DIMENSIONAL GEL DATABASE OF HUMAN CELLULAR PROTEINS

The first step in making a comprehensive 2-dimensional gel protein database is to prepare a synthetic image (digital form of the gel image) of the gel (fluorogram, Coomassie blue or silver stained gel) to be used as a standard or master reference. This can be done with laser scanners, charge couple device (CCD)[2] array scanners, television cameras, rotating drum scanners, and multiwire chambers (13). Computerized analysis systems for spot detection, quantitation, pattern matching, and data handling (access and retrieval of information, database making) have been described in the literature (ELSIE [43], GELLAB [11], HERMeS [44], MELANIE [10], QUEST (9), and TYCHO [8]) and some are available commercially (PDQUEST, Protein Database Inc., Huntington, N.Y.; KEPLER, Large Scale Biology, Rockville, Md.; Visage, BioImage Corporation, Ann Arbor, Mich.; Gemini, Joyce Loebl, Gateshead; Microscan 1000, Technology Resources Inc., Nashville, Tenn. and MasterScan, Billerica, Mass.). Unfortunately, most of these systems are incompatible with one another and their advantages and disadvantages have been discussed by Miller (13).

In our work station in Aarhus, fluorograms are scanned with a Molecular Dynamics laser scanner and the data are analyzed using the PDQUEST II software (Protein Databases Inc.) (12) running on a spark station computer 4100 FC-8-P3 from SUN Microsystems, Inc. The scanner measures intensity in the range of 0–2.0 absorbance. A typical scan of a 17 × 17 cm fluorogram takes about 2 min. Steps in image analysis include: initial smoothing, background substraction, final smoothing, spot detection, and fitting of ideal Gaussian distribution to spot centers. Spot intensity is calculated as the integration of a fitted Gaussian. If calibration strips containing individual segments of a known amount of radioactivity are used, it is possible to merge multiple exposures of the sample image into a single data image of greater dynamic range. Once the synthetic image is created it can be stored on disk and displayed directly on the monitor. Functions that can be used to edit the images include: cancel (for example, to erase scratches that may have been interpreted as spots by the computer; cancel streaks or low dpm spots), combine (sometimes a spot may be resolved into several closely packed spots), restore, uncombine, and add spot to the gel. The process is time consuming—about 1-1/2 day per image. Edited standard images can be matched to other synthetic images. **Figure 2*A*** shows a portion of a standard synthetic image (IEF) of a fluorogram of [^{35}S]methionine labeled cellular proteins from human AMA cells (master database) (20). Images can be displayed either in black and white (resembling the original fluorograms) or in color (other images in Fig. 2), depending on the need. As shown in Fig. 2*B*, each polypeptide is assigned a number by the computer, which facilitates the entry and retrieval of qualitative and quantitative information for any given spot in the gel (20). The standard image can be matched automatically by the computer to other standard or reference gels (Fig. 2*C*, matching of AMA cellular proteins [left] to MRC-5 proteins [right]) provided a few landmark spots are given manually as reference (indicated with a + in Fig. 2*C*) to initiate the process.

---

[2]Abbreviations: CCD, charge couple device; PCNA, proliferating cell nuclear antigen; HPLC, high performance liquid chromatography.

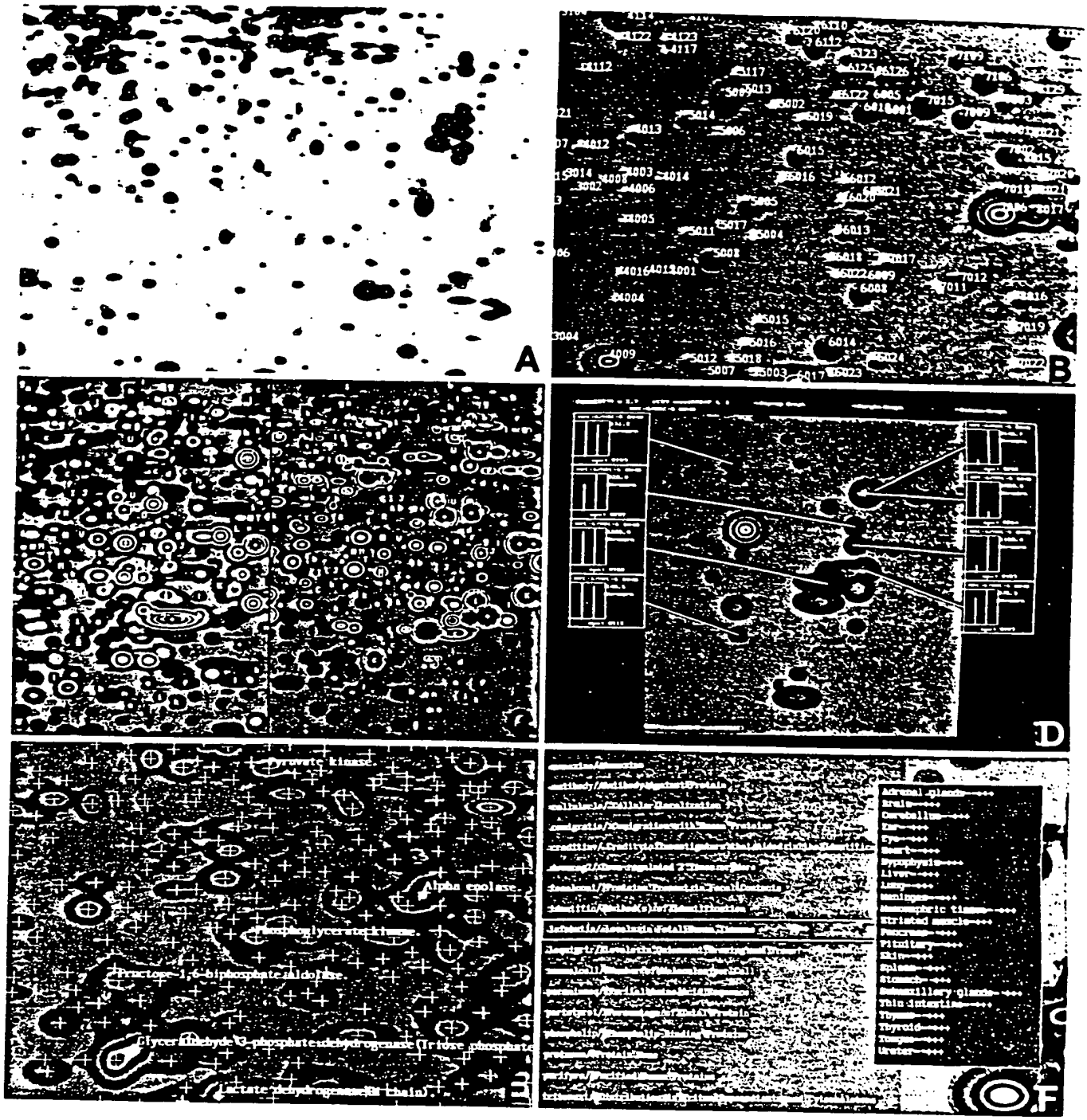
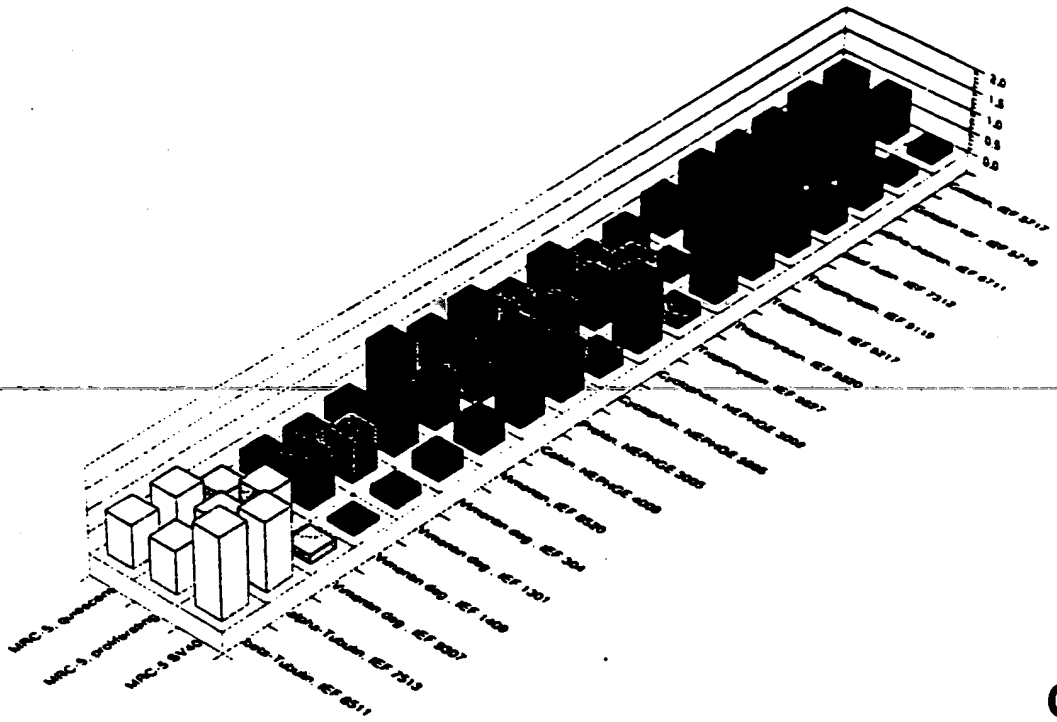
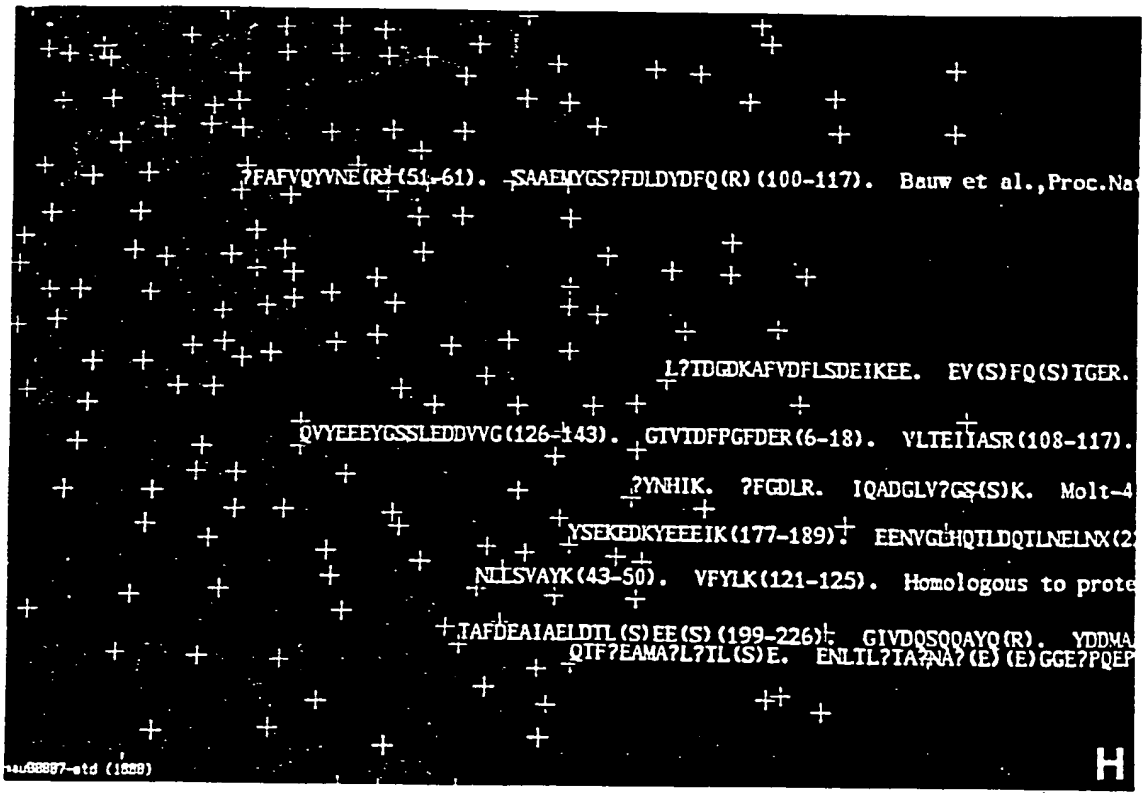


Figure 2. *A*) Synthetic image of a fraction of an IEF gel of the master image of AMA cellular proteins. *B*) As in *A* but showing numbers assigned to each spot. *C*) Comparison of AMA (left) and normal human embryonal lung MRC-5 fibroblasts (right) IEF proteins patterns. Matched proteins are indicated by a + or by the same letters in both gels. Once a protein is matched, information contained in the various categories available in the master AMA database can be transferred. *D*) Synthetic image of a fraction of an IEF fluorogram of [^{35}S]methionine labeled proteins from normal human MRC-5 fibroblasts. The histograms show levels of synthesis of a few proteins in MRC-5 (left bar) and SV40 transformed MRC-5 (right bar) fibroblasts. *E*) Polypeptides that contain information under the category glycolytic pathway. *F*) The function peruse annotation for spot allows the operator to inquire about categories and information available for a given protein. *G*) Relative abundance of cytoskeletal and cytoskeletal-related proteins in quiescent, proliferating, and SV40-transformed MRC-5 fibroblasts. *H*) Polypeptides that contain information under the category partial amino acid sequences.



G



H

The automatic matching process that has been described in detail by Garrels et al. (12) takes about 5 min. Matched proteins are indicated with the same letters in both gels (Fig. 2C). The usefulness of this function is emphasized by the fact that data accumulated on common household proteins can be easily transferred to any other human cellular cell type whose 2-dimensional gel cellular protein pattern is matched to our standard AMA 2-dimensional gel protein image. Alternatively, if the standard gel is part of a matchset (set of gels in a given experiment) it can be used as a linker gel to compare, for example, the quantitative values of a given protein throughout the experiment (see Fig. 2*D*; levels of some proteins in normal and SV40 transformed human MRC-5 fibroblasts) or with other standard images in different sets of

cross-matched experiments (18, 22).

Once a standard map of a given protein sample is made, one can enter qualitative annotations to make a reference database. Our master 2-dimensional gel database of transformed human amnion cell (AMA) proteins (20) lists 3430 polypeptides of which 2592 correspond to cellular components, having pI's ranging from 4 to 13 and molecular weights between 8.5 and 230 kDa. The most abundant proteins in the database correspond to total actin (3.87% of total protein; about 90 million molecules per cell) while the lesser abundant of the recorded polypeptides are present in the vicinity of 5000 molecules per cell. Some annotation categories we are using to establish the master AMA database include: *1*) protein identification (comigration with purified proteins, 2-dimensional immunoblotting, microsequencing); *2*) amounts (total amounts and levels of synthesis); *3*) subcellular localization (nuclear, cytoskeletal, membrane, membrane receptors, specific organelles, etc.); *4*) antibodies; *5*) posttranslational modifications (phosphorylation, glycosylation, methylation etc.); *6*) microsequencing; *7*) cell cycle specificity (specific variations in levels of synthesis and amount); *8*) regulatory behavior (effect of hormones, growth factors, heat shock, etc.) *9*) rate of synthesis in normal and transformed cells (proliferation sensitive proteins, cell cycle specific proteins, oncogenes, components of the pathway (or pathways) that control cell proliferation); *10*) function (mainly from comigration with proteins of known function); *11*) sets of proteins that are coordinately regulated (hierarchy of controls, differential gene expression in various cells, etc.); *12*) cDNAs (cloned cDNAs); *13*) proteins that are specific to a given disease (systematic comparison of protein patterns of fibroblast proteins from healthy and diseased individuals); *14*) expression and exploitation of transfected cDNAs; *15*) pathways (metabolic, others); *16*) gene localization (genetic and physical); *17*) effect of microinjected antibody on patterns of protein synthesis; and *18*) secreted proteins.

Information entered for any spot in a given annotation category can be easily retrieved by asking the computer to display the information on the color screen. For example, Fig. 2*E* shows a synthetic image of a NEPHGE gel (master AMA database) displaying the information contained under the entry glycolytic pathway. Alternatively, one can use the function peruse annotations for spot to directly ask the computer to list all the entries available for a particular protein. By clicking the mouse in a given entry (in this case, presence in fetal human tissues) it is possible to take a quick look at the information in that particular entry (Fig. 2*F*).

A major obstacle encountered in building comprehensive 2-dimensional gel protein databases is identifying the large number of proteins separated by this technology. In our databases (20, 21), known proteins are identified by one or a combination of the following procedures: *1*) comigration with known proteins, *2*) 2-dimensional gel immunoblotting using specific antibodies, and *3*) microsequencing of Coomassie Brillant Blue stained human proteins recovered from dried 2-dimensional gels (see next section). Protein identification by means of microsequencing may be difficult, as individual protein members of families with short peptide differences may escape detection. In the gene-protein database of *E. coli* K-12 (14, 23), another major 2-dimensional gel database available at present, proteins are being identified by a wider range of tests that include comigration with purified proteins; genetic criterion (deletion, insertion, frameshift, nonsense, missense, regulatory), plasmid-bearing strains and in vitro synthesis of protein; selective labeling (methylation, phosphorylation); peptide map similarity; and physiological criterion and selective derivatization.

So far we have received nearly 550 antibodies from laboratories all over the world and these are being systematically tested by 2-dimensional gel immunoblotting for antigen determination. Similarly, purified proteins and organelles provided by several laboratories have greatly aided identification of unknown proteins (20, 21). We routinely request antibodies and protein samples and promise the donors to make available all the information we may have accumulated on that particular protein. For example, **Table 1** lists entries available for Lipocortin V (IEF SSP 8216), also known as annexin V, VAC-α, endonexin II, renocortin, chromobindin-5', anticoagulant protein, PAP-I, γ-calcimedin, IBC, calphobindin, and anchorin CII.

As mentioned previously, one distinct advantage of 2-dimensional gel electrophoresis is the possibility of studying quantitative variations in cellular protein patterns that may lead to identification of groups of proteins that are expressed coordinately during a given biological process. Quantitation, however, is not an easy task as reflected by the lack of published data on global cellular protein patterns. We believe this is partly due to difficulties in obtaining sets of gels that are suitable for computer analysis (streaking, material remaining at the origin, etc.) as well as to limitations (laborious editing time, need of calibration strips to merge images, limited dynamic range, etc.) in the computer analysis systems available at the moment. Perhaps the most advanced quantitative studies published so far using computer analysis have been carried out by Garrels and coworkers (18, 22). In particular, these investigators have established a quantitative rat protein database (18, 22) designed to study growth control (proliferation, growth inhibitors, and stimulation) and transformation in well-defined groups of cell lines obtained by transformation of rat REF52 cells with SV40, adenovirus, and the Kirsten murine sarcoma virus. These studies have revealed clusters of proteins induced or repressed during growth to confluence as well as groups of transformation-sensitive proteins that respond in a differential fashion to transformation by DNA and RNA viruses. A most interesting feature of this quantitative database is the discovery of a group of coregulated proteins that show similar expression patterns as the cell cycle-regulated DNA replication protein known as proliferating cell nuclear antigen (PCNA)/cyclin (45).

In our human databases, most quantitations have been carried out by estimating the radioactivity contained in the polypeptides by direct counting of the gel pieces in a scintillation counter (20, 21). Up to 700 proteins can be cut out through appropriate exposed films in a period of time comparable to that required for editing a synthetic image. Manual quantitation of this large number of spots is difficult without the assistance of a master reference image and a numbering system that can be used to identify the spots. Using this approach, we have recorded quantitative changes in the relative abundance of 592 [^{35}S]methionine-labeled proteins synthesized by quiescent, proliferating, and SV40 transformed human embryonic lung MRC-5 fibroblasts (21). Some data concerning cytoskeletal and cytoskeletal-related proteins are presented in Fig. 2G. Our studies as well as those of Garrels and co-workers (18, 22) may in the long run help define patterns of gene expression that are characteristic of the transformed state.

OTHER 2-DIMENSIONAL GEL PROTEIN DATABASES

As mentioned previously there are other 2-dimensional gel databases available in computer form that have been pub-

TABLE 1. *Some entries for lipocortin V in the human AMA 2-dimensional gel protein database*

Entries for lipocortin V (IEF SSP 8216)	Information entered
1. Protein name	Lipocortin V, renocortin, chromobindin-5, endonexin I, anticoagulant protein, PAP-I, VAC-α, 35-γ-calcimedin, IBC, calphobindin I, anchorin CII, annexin V
2. Percentage of total protein	0.110% (about 2,800,000 molecules per cell)
3. Apparent molecular weight (*mr*)	33.3 kDa
4. Isoelectric point (pI)	4.76
5. Method (or methods) of identification	Microsequencing, 2-dimensional immunoblotting, Comigration
6. Credit to investigators that aided in identification	G. Bauw, J. Vandekerckhove, and colleagues, Rijksuniversiteit Gent; B. Pepinsky, BIOGEN, Cambridge; N.G. Ahn, University of Washington
7. Antibody against protein	Polyclonal (rabbit, antibody no. 20), B. Pepinsky, BIOGEN, Cambridge
8. Comigration with human proteins	Lipocortin V, N.G. Ahn, Howard Hughes Medical Institute, Washington University
9. Cellular localization	Subcortical membrane
10. Calcium/phospholipid-dependent membrane proteins	Lipocortin V
11. Function	Regulation of various aspects of inflammation, immune response, blood coagulation and differentiation
12. Partial amino acid sequence	GTVTDFPGFDER (7–18), VLTEIIASR (109–117), QVYEEEYGSSLEDDVVG (127–143), ?GTDEEKFITIFGT(R) (187–201)
13. cDNA sequence	Known, R. Blake et al., *J. Biol. Chem.* **263**, 10799–10811; 1988 (pI = 4.76 from translated sequence)
14. Levels in fetal human tissues	Adrenal glands = + + +; brain = + + +; cerebellum = + + +; ear = + + +; eye = + + +; heart = + + +; hypophysis = + + +; liver = + + +; lung = + + +; meninges = + + +; mesonephric tissue = + + +; striated muscle = + + +; pancreas = + + +; skin = + + +; spleen = + + +; stomach = + + +; submandibular gland = + + +; small intestine = + + +; thymus = + + +; thyroid gland = + + +; tongue = + + +; ureter = + + +
15. Levels in quiescent, proliferating, and transformed MRC-5 fibroblasts	Q (quiescent) = 1.1; P (proliferating) = 1.0; T (SV40 transformed) = 0.3
16. Distribution in Triton supernatant and cytoskeletons	Mainly supernatant

lished in extenso: these correspond to the *E. coli* K-12 protein-gene database (14, 23) and to the rat REF52 database (18, 22).

The *E. coli* K-12 cellular protein-gene database is perhaps the most complete of all databases reported so far and eventually it should trace each protein back to its structural gene. Information contained in this database includes: gene/protein name (protein name, EC number, gene name); 2-dimensional gel spot designations (x-y coordinates from reference gels, alphanumeric designation); genetic information (linkage map location, physical map location, Genebank code, sequence reference, location on Kohara clones); biochemical information (molecular weight, pI, number of residues of each amino acid, mole percent of each amino acid, total number of amino acids in a polypeptide), and regulatory information (cellular level of protein in different media and different temperature, member of regulon, member of stimulon). Major advances of this database are envisaged in the future in view of the eminent sequencing of the whole *E. coli* genome as well as the development of improved methods to express cloned genes.

The rat REF52 2-dimensional gel protein database lists about 1600 proteins that have been recorded using the QUEST analysis system (18, 22). Included in this quantitative database are *1*) protein names (cytoskeletal and heat shock proteins as well as various nuclear, mitochondrial, and cytoplasmic proteins), *2*) annotations (subcellular localization, modification, recognition by specific antibodies, coprecipitation, NH_2-terminal sequence, cross-reference to protein sequence information and references to the literature), *3*) protein sets (cytoskeletal proteins, phosphoproteins, sets of proteins with PCNA/cyclin-like properties, etc.) and *4*) general quantitative data (protein synthesis during growth of normal REF52 cells to confluence and quiescence, and after restimulation of growth-inhibited cells).

In addition to the 2-dimensional gel databases mentioned so far there are several smaller cellular databases being established in human (normal human diploid fibroblasts, lym-

phocytes, leukocytes, leukemic cells) mouse (NIH/3T3 cells, T lymphocytes), *Aplysia*, yeast (*Saccharomyces cerevisae*), plants (wheat, barley, sorghum), and *Euglena*. Databases of tissue protein, (brain, whole mouse, liver) and body fluid proteins (plasma proteins, cerebrospinal fluid, urine, and milk) are being established in several laboratories. The reader is directed to the review by Celis et al. (4) for details and references concerning these databases.

MICROSEQUENCING HAS ADDED A NEW DIMENSION TO COMPREHENSIVE 2-DIMENSIONAL GEL DATABASES: A DIRECT LINK BETWEEN PROTEINS AND GENES

The development of highly sensitive amino acid gas-phase or liquid-phase sequenators (24), together with the establishment of efficient protein and peptide sample preparation methods, has opened the possibility to perform a systematic sequence analysis of proteins resolved by 2-dimensional gel electrophoresis. Indeed, generated pieces of protein sequences can be used to search for protein identity (comparison with available sequences stored in databanks) as well as for preparing specific DNA probes for cloning of as yet uncharacterized proteins (Fig. 1). In addition, partial protein sequences can be stored in 2-dimensional gel databases (for example, see Fig. 2*H*) and offer a unique link between proteins and genes (Fig. 1).

In the early 1970s gel electrophoresis was used to purify proteins for sequencing purposes (reviewed by Weber and Osborn in ref 25). Proteins were recovered by diffusion and sequenced by the manual dansyl-Edman degradation at the nanomole level. This technique was further refined by using electro-elution to recover proteins and by miniaturizing the system (26). This method has been used extensively, but showed increasing drawbacks (low yields, protein samples contaminated by free amino acids, and NH_2-terminal blocking) as the amounts of handled protein gradually became smaller (e.g., at the 10 picomol level).

Most of the problems referred to above have been minimized with the introduction of protein-electroblotting procedures (27–32). When proteins are blotted on chemically inert membranes, it is possible to sequence the immobilized proteins directly without additional manipulations. Thus, depending on the amount of bound protein and its nature, this direct sequencing procedure generally yields NH_2-terminal sequences containing 10–40 residues. As such, this technique was used to identify, by their NH_2-terminal sequences, differentially expressed major proteins from total cellular extracts separated on 2-dimensional gels. A major difficulty encountered in this procedure is the occurrence of frequent artefactual blockage of the proteins. Several studies suggest that this phenomenon is mainly due to reaction with contaminants (particularly unpolymerized acrylamide present in the gel) and to a high dilution of the protein (low concentration of the protein per unit membrane surface). In addition to this primarily technical problem, many proteins are blocked in vivo by acylation or by a pyrrolidon carboxylic acid cap.

The problem of partial or complete NH_2-terminal blockage can be circumvented by generating internal amino acid sequences. This is achieved by fragmenting the protein present in the gel (gel in situ cleavage) or by cleaving it while bound to the membrane (membrane in situ cleavage) (33–35). In both cases, proteins are either cleaved in a restricted way (e.g., by limited enzymatic digestion or by using restriction chemical cleavage conditions) or fragmented into smaller peptides.

Of the different combinations examined, we had good results by using exhaustive proteolytic digestion on membrane-immobilized proteins. This method has been described for Ponceau red-stained proteins on nitrocellulose blots (34), for Amido-black-stained Immobilon-bound proteins, and for fluorescamine-detected proteins on glass fiber membranes (35). The proteases used (trypsin, chymotrypsin, or pepsin) cleave at multiple sites, generating small peptides that elute from the blot into the digestion buffer from which they are purified by reversed-phase high performance liquid chromatography (HPLC) before being sequenced individually. Although each of these manipulations could be expected to result in a reduced yield of final sequence information, we were surprised that the peptides could be sequenced with high efficiency. In our hands, this approach could be routinely applied to gel-purified proteins available in amounts ranging from 5 to 10 μg, and often yielded sequence information covering more than 30% of the total protein. As membrane-immobilized proteins are not homogeneously digested, but rather show protease sensitivity next to resistant regions, the number of peptides generated is much lower than expected from the number of potential cleavage sites. Consequently, HPLC peptide chromatograms are less complex and most peptides can be recovered in pure form.

As only limited amounts of a protein mixture can be loaded on a 2-dimensional gel, proteins of interest are often obtained in yields insufficient for the currently available sequencing technology. More material can be obtained by enriching for a certain subcellular fraction (purified cell organelles) or by exploiting affinity (dyes, metals, drugs, etc) or hydrophobic properties of proteins before gel analysis. All of the sequencing results accumulated so far in the human protein database (20) (a few are shown in Fig. 2*H*) have been obtained from analysis of protein spots collected from 2-dimensional gels that had been stained with Coomassie blue according to standard procedures and dried for storage. Proteins are recovered from the collected gel pieces by a protein-elution-concentration device, combined with gel electrophoresis and electroblotting. Details of this technique have been reported in a previous communication (42) and a brief outline is given below.

Combined gel pieces are allowed to swell in gel sample buffer (a total volume of 1.5 ml). The gel pieces combined with the supernatant are then collected into a large slot made in a new gel. The slot is further filled with Sephadex G-10 equilibrated in gel sample buffer. During consecutive gel electrophoresis, most of the electrical current passes on the side of the slot instead of passing through the slot. This results in both a vertical stacking and horizontal contraction of the protein band. With this device the protein is efficiently eluted from the gel pieces and concentrated from a large volume into a narrow spot. The highly concentrated (about 5 mm^2) protein spot is then electroblotted on PVDF-membranes, stained with Amido black, and in situ digested with trypsin. The peptides generated during digestion elute from the membrane into the supernatant, and can be separated by narrow bore reversed-phase HPLC and collected individually for sequence analysis.

Using this and previous procedures (37, 39, 42), we have so far analyzed 70 protein spots collected from 2-dimensional gels (20, and unpublished observations) (see for example Fig. 2*H*). The sequence information amounts to 2100 allocated residues corresponding to an average of 30 residues per protein spot. So far we have made cDNAs of many of the unknown proteins that have been microsequenced, and a substantial number has been cloned and sequenced. All available information indicates that it may be possible to obtain partial sequence information from most of

the proteins that can be visualized by Coomassie Brillant Blue staining.

Partial protein sequences are stored in the database as displayed in Fig. 2*H*, and it should be possible in the near future to interface this information with forthcoming DNA sequence data from the human genome project. In the long run, as the human genome sequences become available it will be possible to assign partial protein sequences to genes for which the full DNA sequence and chromosomal location are known (Fig. 1).

SUMMARY

The studies presented in this brief review are intended to demonstrate the usefulness of computer-aided 2-dimensional gel electrophoresis and microsequencing to analyze cellular protein patterns, and to link protein and DNA information. As more information is gathered worldwide, comprehensive databases will depict an integrated picture of the expression levels and properties of the thousands of proteins that orchestrate most cellular functions.

Clearly, databases allow easy access to a large body of data and provide an efficient medium to communicate standardized protein information. In the future, databases will foster a wide variety of biological information that can be used to support collaborative research projects in basic and applied biology as well as in clinical research (2, 5, 46). Once a protein is identified in a particular database all the information gathered on it can be made available to the scientist. However, many problems must be solved before protein databases become of general use to the scientific community. A most urgent one is to promote standardization of the gel running conditions so that data produced in a given laboratory may be used worldwide. Surprisingly, the gel running technology as it stands today is still a craftmanship art.

Finally, comprehensive, computerized databases of proteins, together with recently developed techniques to microsequence proteins, offer a new dimension to the study of genome organization and function (Fig. 1). In particular, human protein databases may become increasingly important in view of the concerted effort to map and sequence the entire human genome. This formidable task is expected to dominate biological research in the next decades. FJ

We would like to thank S. Himmelstrup Jørgensen for typing the manuscript and O. Sønderskov for photography. Work in the authors' laboratories was supported by grants from the Danish Biotechnology Programme, the Danish Cancer Foundation, and the Commission of the European Communities.

REFERENCES

1. O'Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021
2. Special Issue: Two-dimensional gel electrophoresis. *Clin. Chem.* **28**, 1982
3. Celis, J. E., and Bravo, R., eds. (1984) *Two-Dimensional Gel Electrophoresis of Proteins: Methods and Applications*. Academic, New York
4. Celis, J. E., Madsen, P., Gesser, B., Kwee, S., Nielsen, H. V., Rasmussen, H. H., Honoré, B., Leffers, H., Ratz, G. P., Basse, B., Lauridsen, J. B., and Celis, A. (1989) Protein databases derived from the analysis of two-dimensional gels. In *Advances in Electrophoresis* (Chrambach, C., ed) VCH, Weinheim, Germany
5. Special Issue: Two-dimensional gel electrophoresis in cell biology. (Celis, J. E., ed) *Electrophoresis* **11**, 1990
6. Celis, J. E., Honoré, B., Bauw, G., and Vandekerckhove, J. (1990) Comprehensive computerized 2D gel protein databases offer a global approach to the study of the mammalian cell. *BioEssays* **12**, 93–98
7. Garrels, J. I. (1983) Two-dimensional gel electrophoresis and computer analysis of proteins synthesized by cloned cell lines. *Methods Enzymol.* **100**, 411–423
8. Anderson, N. L., Hofmann, J. P., Gemmel, A., and Taylor, S. (1984) Global approaches to the quantitative analysis of gene-expression patterns observed by two-dimensional gel electrophoresis. *Clin. Chem.* **30**, 2031–2036
9. Garrels, J. I., Farrar, J. T., and Burwell, C. B. (1984) The Quest system for computer-analyzed two-dimensional electrophoresis of proteins in *Two-Dimensional Gel Electrophoresis of Proteins. Methods and Applications* (Celis, J. E., and Bravo, R., eds) pp. 37–91, Academic, New York
10. Vincens, P., and Tarroux, P. (1988) Two-dimensional electrophoresis computerized processing. *Int. J. Biochem.* **20**, 499–509
11. Appel, R., Hochstrasser, D., Roch, C., Funk, M., Muller, A. F., and Pellegrini, C. (1988) Automatic classification of two-dimensional gel electrophoresis pictures by heuristic clustering analysis: a step toward machine learning. *Electrophoresis* **9**, 136–142
12. Lemkin, P. F., and Lester, E. P. (1989) Database and search techniques for two-dimensional gel protein data: a comparison of paradigms for exploratory data analysis and prospects for biological modeling. *Electrophoresis* **10**, 122–140
13. Miller, M. J. (1989) Computer-assisted analysis of two-dimensional gel electrophoretograms. *Adv. Electrophoresis* **3**, 182–217
14. Phillips, T. D., Vaughn, V., Bloch, P. L., and Neidhardt, F. C. (1987) In *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology, Gene-Protein Index of Escherichia coli K-12*, 2 ed. (Neidhardt, F. C., Ingraham, J. I., Low, K. B., Magasanik, B., Schaechter, M., and Umbarger, H. E. ed) pp. 919–966, American Society for Microbiology, Washington, D.C.
15. Celis, J. E., Ratz, G. P., Celis, A., Madsen, P., Gesser, B., Kwee, S., Madsen, P. S., Nielsen, H. V., Yde, H., Lauridsen, J. B., and Basse, B. (1988) Towards establishing comprehensive databases of cellular proteins from transformed human epithelial amnion cells (AMA) and normal peripheral blood mononuclear cells. *Leukemia* **9**, 561–601
16. Special Issue: Protein databases in two-dimensional electrophoresis. (Celis, J. E., ed) *Electrophoresis* **2**, 1989
17. Celis, J. E., Ratz, G. P., Madsen, P., Gesser, B., Lauridsen, J. B., Brogaard-Hansen, K. P., Kwee, S., Rasmussen, H. H., Nielsen, H. V., Crüger, D., Basse, B., Leffers, H., Honoré, B., Møller, O., and Celis, A. (1989) Computerized, comprehensive databases of cellular and secreted proteins from normal human embryonic lung MRC-5 fibroblasts: identification of transformation and/or proliferation sensitive proteins. *Electrophoresis* **10**, 76–115
18. Garrels, J. I., and Franza, B. R. (1989) The REF52 protein database. Methods of database construction and analysis using the Quest system and characterizations of protein patterns from proliferating and quiescent REF52 cells. *J. Biol. Chem.* **264**, 5283–5298
19. Celis, J. E., Crüger, D., Kiil, J., Dejgaard, K., Lauridsen, J. B., Ratz, G. P., Basse, B., Celis, A., Rasmussen, H. H., Bauw, G., and Vandekerckhove, J. (1990) A two-dimensional gel protein database of noncultured total normal human epidermal keratinocytes: identification of proteins strongly up-regulated in psoriatic epidermis. *Electrophoresis* **11**, 242–254
20. Celis, J. E., Gesser, B., Rasmussen, H. H., Madsen, P., Leffers, H., Dejgaard, K., Honoré, B., Olsen, E., Ratz, G., Lauridsen, J. B., Basse, B., Mouritzen, S., Hellerup, M., Andersen, A., Walbum, E., Celis, A., Bauw, G., Puype, M., Van Damme, J., and Vandekerckhove, J. (1990) Comprehensive two-dimensional gel protein databases offer a global approach to the analysis of human cells: the transformed amnion cells (AMA) master database and its link to genome DNA sequence data. *Electrophoresis* **12**, 989–1071

21. Celis, J. E., Dejgaard, K., Madsen, P., Leffers, H., Gesser, B., Honoré, B., Rasmussen, H. H., Olsen, E., Lauridsen, J. B., Ratz, G., Mouritzen, S., Hellerup, M., Andersen, A., Walbum, E., Celis, A., Bauw, G., Puype, M., Van Damme, J., and Vandekerckhove, J. (1990) The MRC-5 human embryonal lung fibroblast two-dimensional gel cellular protein database: quantitative identification of polypeptides whose relative abundance differs between quiescent, proliferating and SV40 transformed cells. *Electrophoresis* **12**, 1072–1113
22. Garrels, J. I., Franza, B. R., Chang, C., and Latter, G. (1990) Quantitative exploration of the REF52 protein database: cluster analysis reveals the major protein expression profiles in responses to growth regulation, serum stimulation, and viral transformation. *Electrophoresis* **12**, 1114–1130
23. Van Bogelen, R. A., Hutton, M. E., and Neidhardt, F. C. (1990) Gene-protein database of *Escherichia coli* K-12, 3rd ed. *Electrophoresis* **12**, 1131–1166
24. Hewick, R. M., Hunkapiller, M. W., Hood, L. E., and Dreyer, W. J. (1981) A gas-liquid solid phase peptide and protein sequenator. *J. Biol. Chem.* **256**, 7990–7997
25. Weber, K., and Osborn, M. (1985) In *The Proteins and Sodium Dodecyl Sulfate: Molecular Weight Determination on Polyacrylamide Gels and Related Procedures* (Neurath, H. et al., eds) Vol. 1, pp. 179–223, Academic, New York
26. Hunkapiller, M. W., Lujan, E., Ostrander, F., and Hood, L. E. (1983) Isolation of microgram quantities of proteins from polyacrylamide gels for amino acid sequence analysis. *Methods Enzymol.* **91**, 227–236
27. Vandekerckhove, J., Bauw, G., Puype, M., Van Damme, J., and Van Montagu, M. (1985) Protein-blotting on polybrene-coated glass-fiber sheets. *Eur. J. Biochem.* **152**, 9–19
28. Aebersold, R. H., Teplow, D. B., Hood, L. E., and Kent, S. B. H. (1986) Electroblotting onto activated glass. *J. Biol. Chem.* **261**, 4229–4238
29. Bauw, G., De Loose, M., Inzé, D., Van Montagu, M., and Vandekerckhove, J. (1987) Alterations in the phenotype of plant cells studied by NH_2-terminal amino acid-sequence analysis of proteins electroblotted from two-dimensional gel-separated total extracts. *Proc. Natl. Acad. Sci. USA* **84**, 4806–4810
30. Matsudaira, P. (1987) Sequence from picomole quantities of proteins electroblotted onto polyvinylidene difluoride membranes. *J. Biol. Chem.* **262**, 10035–10038
31. Eckerskorn, C., Mewes, W., Goretzki, H., and Lottspeich, F. (1985) A new siliconized-glass fiber as support for protein-chemical analysis of electroblotted proteins. *Eur. J. Biochem.* **176**, 509–519
32. Moos, M., Jr., Nguyen, N. Y., and Liu, T.-Y. (1988) Reproducible high yield sequencing of proteins electrophoretically separated and transferred to an inert support. *J. Biol. Chem.* **263**, 6005–6008
33. Kennedy, T. E., Gawinowicz, M. A., Barzilai, A., Kandel, E. R., and Sweatt, J. D. (1988) Sequencing of proteins from two-dimensional gels by using in situ digestion and transfer of peptides to polyvinylidene difluoride membranes: application to protein associated with sensitization in *Aplysia*. *Proc. Natl. Acad. Sci. USA* **85**, 7008–7012
34. Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E., and Kent, S. B. H. (1987) Internal amino acid sequence analysis of protein separated by one- or two-dimensional gel electrophoresis after in situ protease digestion on nitrocellulose. *Proc. Natl. Acad. Sci. USA* **84**, 6970–6974
35. Bauw, G., Van Den Bulcke, M., Van Damme, J., Puype, M., Van Montagu, M., and Vandekerckhove, J. (1988) Protein electroblotting on polybase-coated glassfiber and polyvinylidine difluoride membranes: an evaluation. *J. Prot. Chem.* **7**, 194–196
36. Celis, J. E., Ratz, G. P., Madsen, P., Gesser, B., Lauridsen, J. B., Leffers, H., Rasmussen, H. H., Nielsen, H. V., Crüger, D., Basse, B., Honoré, B., Møller, O., Celis, A., Vandekerckhove, J., Bauw, G., Van Damme, J., Puype, M., and Van Den Bulcke, M. (1989) Comprehensive, human cellular protein databases and their implication for the study of genome organization and function. *FEBS Lett.* **244**, 247–254
37. Bauw, G., Van Damme, J., Puype, M., Vandekerckhove, J., Gesser, B., Lauridsen, J. B., Ratz, G. P., and Celis, J. E. (1989) Protein-electroblotting and -microsequencing strategies in generating protein databases from two-dimensional gels. *Proc. Natl. Acad. Sci. USA* **86**, 7701–7705
38. Aebersold, R., and Leavitt, J. (1990) Sequence analysis of proteins separated by polyacrylamide gel electrophoresis. Towards an integrated protein database. *Electrophoresis* **11**, 517–527
39. Bauw, G., Rasmussen, H. H., Van Den Bulcke, M., Van Damme, J., Puype, M., Gesser, B., Celis, J. E., and Vandekerckhove, J. (1990) Two-dimensional gel electrophoresis, protein electroblotting and microsequencing: a direct link between proteins and genes. *Electrophoresis* **11**, 528–536
40. Tempst, P., Link, A. J., Riviere, L. R., Fleming, M., and Elicone, C. (1990) Internal sequence analysis of protein separated on polyacrylamide gels at the submicrogram level: improved methods, applications and gene cloning strategies. *Electrophoresis* **11**, 537–553
41. Eckerskorn, C., and Lotspeich, F. (1990) Combination of two-dimensional gel electrophoresis with microsequence and amino acid composition analysis: improvement of speed and sensitivity in protein characterization. *Electrophoresis* **11**, 554–561
42. Rasmussen, H. H., Van Damme, J., Bauw, G., Puype, M., Gesser, B., Celis, J. E., and Vandekerckhove, J. (1991) In *Methods in Protein Sequence Analysis* (Jörnvall, H., and Höög, J. O., eds) pp. 103–114, Eighth International Conference on Methods in Protein Sequence Analysis, Birkhäuser Verlag, Boston
43. Olson, A. D., and Miller, M. J. (1988) Elsie 4: quantitative computer analysis of sets of two-dimensional gel electrophoretograms. *Anal. Biochem.* **169**, 49–70
44. Vincens, P., Paris, N., Pujol, J. L., Gaboriaud, C., Rabilloud, T., Pennetier, J., Matherat, P., and Tarroux, P. (1986) HERMeS: a second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part I: Data acquisition. *Electrophoresis* **7**, 347–356
45. Celis, J. E., Madsen, P., Celis, A., Nielsen, H. V., and Gesser, B. (1987) Cyclin (PCNA, auxiliary protein of DNA polymerase-δ) is a central component of the pathway(s) leading to DNA replication and cell division. *FEBS Lett* **220**, 1–7
46. Anderson, N. G., and Anderson, N. L. (1982) The human protein index. *Clin. Chem.* **28**, 739–748